A compute-in-memory chip based on resistive random-access memory

Year: 2022 | Citations: 636 | Authors: W. Wan, R. Kubendran, Clemens J. S. Schaefer, S. Eryilmaz, Wenqiang Zhang

Abstract

Realizing increasingly complex artificial intelligence (AI) functionalities directly on edge devices calls for unprecedented energy efficiency of edge hardware. Compute-in-memory (CIM) based on resistive random-access memory (RRAM)1 promises to meet such demand by storing AI model weights in dense, analogue and non-volatile RRAM devices, and by performing AI computation directly within RRAM, thus eliminating power-hungry data movement between separate compute and memory2–5. Although recent studies have demonstrated in-memory matrix-vector multiplication on fully integrated RRAM-CIM hardware6–17, it remains a goal for a RRAM-CIM chip to simultaneously deliver high energy efficiency, versatility to support diverse models and software-comparable accuracy. Although efficiency, versatility and accuracy are all indispensable for broad adoption of the technology, the inter-related trade-offs among them cannot be addressed by isolated improvements on any single abstraction level of the design. Here, by co-optimizing across all hierarchies of the design from algorithms and architecture to circuits and devices, we present NeuRRAM—a RRAM-based CIM chip that simultaneously delivers versatility in reconfiguring CIM cores for diverse model architectures, energy efficiency that is two-times better than previous state-of-the-art RRAM-CIM chips across various computational bit-precisions, and inference accuracy comparable to software models quantized to four-bit weights across various AI tasks, including accuracy of 99.0 percent on MNIST18 and 85.7 percent on CIFAR-1019 image classification, 84.7-percent accuracy on Google speech command recognition20, and a 70-percent reduction in image-reconstruction error on a Bayesian image-recovery task. A compute-in-memory neural-network inference accelerator based on resistive random-access memory simultaneously improves energy efficiency, flexibility and accuracy compared with existing hardware by co-optimizing across all hierarchies of the design.